# Structural Divergence under Shared Language-Level Specification: Griko in Universal Dependencies

**Stavros Bompolas[*1], Emanuela Pinna[*2], Josep Quer[2], Marika Lekakou[3], Stella Markantonatou[4]**

[1]Archimedes/Athena RC, [2]Universitat Pompeu Fabra, [3]University of Ioannina,
[4]Institute of Language and Speech Processing/Athena RC
[1]Artemidos 1, 15125, Marousi, Greece, [2]Carrer de Roc Boronat 138, 08018, Barcelona, Spain,
[3]Department of Philology, 45110, Ioannina, Greece, [4]Aigialias 19 & Chalepa 3, 15125, Marousi, Greece
{s.bompolas, marks}@athenarc.gr, {josep.quer, emanuela.pinna}@upf.edu, mlekakou@uoi.gr

**Abstract**

Dialectal varieties pose major challenges for NLP resource development, especially when annotation frameworks are organized around standardized language specifications. In Universal Dependencies (UD), dialects without independent ISO codes are subsumed under the corresponding standard language and inherit its language-level documentation, validator settings, and grammatical inventories. This paper examines Griko, a Greek variety spoken in southern Italy that developed in relative isolation from the Modern Greek dialect continuum while remaining in long-term contact with local Italo-Romance varieties. We assess the consequences of this organizational structure through controlled parsing experiments comparing intra-dialectal training, cross-dialectal transfer from Standard Modern Greek (SMG), script-controlled transfer using romanized SMG, and contact-related cross-lingual transfer from Italian. Our results show that, before romanization, the Italian model even surpasses SMG on several UD metrics and that, although romanization substantially improves SMG-based transfer, performance still remains far below the intra-dialectal baseline. We argue that this persistent gap reflects the interaction between structural divergence and language-level validation constraints, a phenomenon we term *ISO-based validation coupling*. Through analyses of auxiliary systems, voice marking, and progressive constructions, we show how standard-centric validation architectures can constrain the representation of dialect-specific grammar. More broadly, the Griko case highlights the limitations of language-centric organization in UD and underscores the need for variety-sensitive mechanisms when extending universal annotation frameworks to structurally divergent dialects.

**Keywords:** Universal Dependencies, dialectal NLP, Griko, cross-dialectal transfer, morphosyntactic divergence

## 1. Introduction

Dialectal varieties pose fundamental challenges for NLP resource development, particularly when annotation frameworks are designed around standardized language varieties. Within the Universal Dependencies (UD) framework (Nivre et al., 2020; de Marneffe et al., 2021), dialects that lack independent ISO codes are typically subsumed under the standard language, implicitly assuming structural comparability and transferability. As a result, annotation decisions are constrained by shared inventories, validation rules, and morphosyntactic specifications defined at the language level.

We present a case study of Griko (ISO 639-3: `ell`; Glottolog: `apul1237`), a Modern Greek (MG) variety spoken in Salento (southern Italy). Griko has evolved in relative isolation from the MG dialect continuum and in prolonged contact with the local Italo-Romance varieties, resulting in extensive phonological and morphosyntactic divergence. Because Griko lacks its own ISO code, it is represented under the MG specification in UD, alongside Standard Modern Greek (SMG) and other MG varieties. As a consequence, dialect-specific grammatical categories cannot be independently specified or constrained without potentially affecting all Greek varieties grouped under the same language code, and vice versa. This organizational constraint has direct implications for annotation design, validation procedures, and computational modeling.

Griko has recently attracted attention in NLP as a severely under-resourced and endangered variety. Existing work has developed parallel corpora, POS-tagging resources, and speech processing datasets (Boito et al., 2018; Anastasopoulos et al., 2018; Vakirtzian et al., 2024), as well as alignment and correction methods for noisy or low-resource data (Rijhwani et al., 2020; Xie and Anastasopoulos, 2023). While these studies demonstrate the feasibility of applying NLP techniques to Griko (see also Ramponi, 2024), they primarily address task-specific modeling rather than foundational questions of syntactic resource design within a universal annotation framework.

In this paper, we address this gap by investigating the morphosyntactic and computational con-

*Equal contribution.



*Proceedings of Workshop on Dialects in NLP — A Resource Perspective (DialRes)* , pages 315–326
16 May 2026. 

sequences of incorporating Griko into UD under the MG specification. We argue that this standard-centric treatment creates both technical and theoretical challenges. To substantiate this claim, we evaluate four training configurations: (i) intra-dialectal training (Griko→Griko); (ii) cross-dialectal transfer from Standard Modern Greek (SMG→Griko); (iii) script-controlled cross-dialectal transfer using romanized SMG data (Romanized SMG→Griko); and (iv) contact-related cross-lingual transfer from Italian (Italian→Griko). The results show that Italian slightly outperforms non-romanized SMG on some UD metrics, suggesting that script alignment and contact-related overlap may offer limited advantages, although both transfer conditions remain weak overall. Crucially, even after orthographic differences are minimized, transfer from SMG to Griko remains extremely poor. This indicates that script discrepancies alone do not explain transfer failure; deeper morphosyntactic divergence, together with annotation-level constraints, plays a decisive role. Therefore, we examine specific annotation challenges, including auxiliary inventories, voice distinctions, and progressive constructions, to show how the shared UD specification obscures structural mismatches between Griko and SMG.

Our findings suggest that neither genealogical relatedness nor shared ISO coding is sufficient to guarantee structural compatibility for dependency parsing. Taken together, these results motivate a broader critique of UD as a language-centric framework. More broadly, we argue that standard-centric modeling practices and language-level validation architectures can obscure dialect-specific grammatical systems, thereby affecting both computational performance and faithful linguistic representation.

By combining empirical parsing experiments with annotation-level analysis, this study contributes to ongoing discussions on how universal frameworks, such as UD, can accommodate dialectal diversity without enforcing structural homogenization.

## 2. Background

This section situates the present study within the broader landscape of dialectal resource development in UD. We first outline how dialects are currently represented within the UD ecosystem and discuss emerging efforts to model non-standard varieties. We then review the status of MG in UD and examine how its organizational and validation structure affects the inclusion of additional varieties. This overview provides the necessary context for understanding the specific challenges posed by Griko.

### 2.1. Dialects in Universal Dependencies

The UD framework has made substantial progress in multilingual morphosyntactic annotation, currently covering more than 150 languages. Its cross-linguistically consistent design has enabled large-scale comparative research and transfer-based NLP modeling. However, extending UD to dialectal varieties exposes important theoretical and methodological challenges.

Recent work has increasingly focused on developing dialectal treebanks. According to a recent survey,[1] more than 30 dialectal UD treebanks have been identified. Treebanks have been created for dialects and non-standard varieties such as Egyptian Arabic (Maamouri et al., 2014), Norwegian dialects (Øvrelid et al., 2018; Kåsen et al., 2022), Occitan (Miletic et al., 2020), and Bavarian (Blaschke et al., 2024), among others. These initiatives reflect a growing recognition that dialectal resources are essential both for linguistic documentation and for building robust NLP systems capable of handling non-standard input.

Despite this progress, dialectal resources in UD remain constrained by the framework's language-based organization. UD's higher-level metadata are explicitly language-centered: the current language inventory states that information about language families and genera is mostly taken from WALS Online (Dryer and Haspelmath, 2013), a resource well suited to broad genealogical classification but too coarse to encode dialect-internal structural divergence. As a result, varieties without distinct ISO codes are often subsumed under a standard-language node and inherit language-level documentation and validation assumptions, including inventories of auxiliaries, morphological features, and consistency constraints. At the same time, UD also includes non-standard varieties that are treated as separate language entries with their own documentation, including cases that are neither separately represented in WALS nor assigned a distinct ISO code, such as Pomak (Markantonatou et al., 2023). The coexistence of these organizational strategies creates a structural tension between cross-dialectal comparability and dialect-specific adequacy, raising broader questions about how grammatical divergence can be represented within a shared language specification.

### 2.2. Modern Greek and its dialects in Universal Dependencies

The UD framework has been applied to SMG through two major treebanks: UD_Greek-GDT

[1] https://unidive.lisn.upsaclay.fr/lib/exe/fetch.php?media=meetings:wg1:worshop-submission-unidive.pdf

(Prokopidis and Papageorgiou, 2017), based on the Greek Dependency Treebank, and the more recent UD_Greek-GUD (Markantonatou et al., 2025), which adheres more closely to UD.v2 morphological guidelines. The two treebanks differ with respect to the registers they cover and the detail of modeling linguistic phenomena.

These resources serve as benchmarks for dialect-oriented NLP research. Newly developed treebanks for MG varieties are generally expected to align with SMG-based annotation guidelines to ensure compatibility within the UD ecosystem.

In recent years, five UD treebanks have been developed for MG dialects. Two focus on Cappadocian—UD_Cappadocian-AMGiC (Sampanis and Prokopidis, 2021) and UD_Cappadocian-TueCL (Vligouridou et al., 2024)—while three represent other dialects: UD_Greek-Cretan (Vakirtzian et al., 2025), UD_Greek-Lesbian (Bompolas et al., 2025), and UD_Greek-Messenian.

Cretan, Lesbian, and Messenian are annotated under the shared MG ISO code and follow the SMG-based morphosyntactic specification, maintaining substantial structural continuity with the standard variety.

Cappadocian, however, constitutes a different case. Like Griko, Cappadocian developed over centuries in relative isolation and under intense contact influence—primarily from Turkish (Dawkins, 1916). Crucially, Cappadocian is assigned its own ISO 639-3 code (`cpg`) and is therefore treated as a separate language in UD. Its treebanks are not bound by the MG validation constraints and can specify independent auxiliary inventories and feature systems without affecting other Greek treebanks.

Griko represents a structurally comparable case of long-term isolation and contact-induced change, yet it does not possess a separate ISO code. As a result, it must be incorporated under the MG language specification. This means that its annotation is subject to the same validation rules and feature inventories as SMG. Consequently, dialect-specific grammatical distinctions cannot be formalized without affecting all treebanks grouped under the same language code.

Incorporating Griko into UD is, therefore, not simply a matter of adding a new dataset. It requires reconciling dialectal morphosyntactic divergence with a validation architecture designed for the standard language.

## 2.3. The Griko Variety

Griko is a Greek variety spoken in Salento, in the south-eastern extremity of Italy, and has been attested at least since the Middle Ages. Its precise origin has long been debated. The main hypotheses summarized in Manolessou (2005, and references therein) converge into three principal theories: (a) Griko reflects an unbroken Greek presence in Southern Italy dating back to the Magna Graecia colonies of the 8th–7th centuries BCE; (b) it originated during the Byzantine presence in Southern Italy between the 6th and 11th centuries CE; or (c) like most MG dialects (with the exception of Tsakonian), it participated in the broader evolution from Hellenistic Koiné to Late Medieval Greek through continuous contact between Southern Italy and the Greek-speaking world, with archaic traits reflecting preserved Koiné features. Today, theories (b) and (c) are generally regarded as the most plausible.

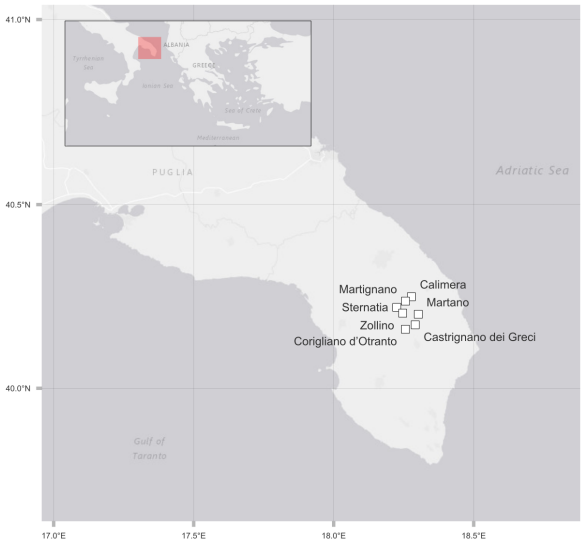


Figure 1: Geographic distribution of the remaining Griko-speaking communities, concentrated in seven towns of Salento (Apulia, Southern Italy).

Regardless of its precise historical origin, Griko has developed for centuries in sustained contact with Salentino, the local Italo-Romance variety, and later with Standard Italian. This prolonged bilingual setting has resulted in extensive contact-induced change across phonology, lexicon, and morphosyntax. Several core grammatical domains—including auxiliary formation, voice marking, and progressive constructions—show restructuring aligned with Italo-Romance patterns. These developments are particularly relevant for syntactic annotation, as they diverge systematically from SMG.

From a demographic perspective, Griko is classified by UNESCO as a severely endangered language (Moseley, 2012). Intergenerational transmission has largely ceased since the Second World War, and the language is today mainly spoken by elderly speakers. Its remaining speech communities are concentrated in seven towns in Salento (Figure 1)—Calimera, Castrignano dei Greci, Corigliano d'Otranto, Martano, Martignano, Sternatia, and Zollino—where vitality varies. Sociolinguistically, Griko has historically existed in a diglossic relation-

ship with Salentino and Italian, functioning as the lower-prestige code largely restricted to informal and domestic domains (see Pellegrino, 2021 for a comprehensive diachronic and synchronic account).

Griko has traditionally been transmitted orally and still lacks an officially standardized orthography. Written documentation, primarily in the form of poetry and narrative texts, began to appear in the late 19th century and consistently employs a Latin-based alphabet. The absence of orthographic standardization introduces additional challenges for corpus development and NLP processing, especially when resources are integrated into infrastructures primarily designed for Greek-script varieties.

Taken together, Griko represents a case of long-term contact-driven divergence within the MG dialect continuum. Its structural development under conditions of isolation, bilingualism, and sociolinguistic marginalization renders it particularly informative for examining how standard-centric annotation frameworks accommodate—or fail to accommodate—deeply divergent dialectal systems.

## 3. Experimental Design

The issues discussed above motivate an empirical evaluation of transfer-based dependency parsing within UD. To this end, we conduct controlled parsing experiments comparing intra-dialectal, cross-dialectal, script-controlled, and contact-related cross-lingual training conditions. This design allows us to test whether shared ISO-based grouping actually yields cross-variety parsing compatibility, or whether morphosyntactic divergence between Griko and SMG constrains transfer despite genealogical relatedness.

### 3.1. Data

**Griko Dataset.** The Griko data used in this study originate from the Palumbo corpus (Tommasi, 1998), a collection of 114 narrative texts originally recorded and documented by Vito Domenico Palumbo (1854–1928). The corpus was recently digitized and enriched with Italian translations and partial gold part-of-speech annotation (Anastasopoulos et al., 2018).

For the present study, a subset of 639 sentences from the Palumbo corpus was manually annotated and converted into CoNLL-U format by a trained linguist with specialized expertise in Griko grammar. Annotation decisions followed the UD.v2 guidelines and were aligned, where structurally possible, with the conventions adopted in the UD_Greek-GUD treebank. However, full alignment with SMG was not always feasible due to systematic morphosyntactic divergence. In such cases, annotation choices prioritized preserving grammatical distinctions internal to Griko.

The Griko treebank should currently be considered under development. For the experiments reported here, validation was restricted to format-level compliance using the official UD CoNLL-U validator.[2] Language-specific morphological and syntactic validation rules were not fully enforced, as these are presently defined only for SMG under the shared ISO code.

The dataset was partitioned into training and test splits; for the trainable conditions, Stanza automatically created the development split from the training portion.

**Standard Modern Greek Dataset.** For SMG, we use the UD_Greek-GUD treebank (Markantonatou et al., 2025). UD_Greek-GUD (GUD) is the most recent UD treebank for SMG. The corpus comprises fiction texts as well as material from online sources that reflect colloquial SMG.

This dataset represents the standard-based morphosyntactic specification to which additional MG varieties must formally align within the UD framework. It therefore serves as the baseline source for the cross-dialectal transfer experiments reported in this study. However, SMG is used here primarily as a convenient reference variety and comparison baseline, rather than as the most historically appropriate Greek point of comparison for Griko.

For the purposes of this study, we used the most current version of the GUD training set. In trainable conditions, Stanza automatically derived the development split from the training data.

**Romanized SMG Dataset (Script-Control Condition).** To isolate the effect of script differences between SMG (Greek alphabet) and Griko (Latin alphabet), we created a romanized version of the GUD treebank.

Romanization was performed automatically using a rule-based transliteration pipeline (uroman; Hermjakob et al., 2018)[3] applied directly to the CoNLL-U files. The procedure preserves tokenization, morphological features, and dependency structure while transforming only the orthographic representation into Latin script.

This dataset allows us to evaluate whether transfer degradation from SMG to Griko is primarily due to script mismatch or instead reflects deeper structural divergence.

[2] https://github.com/UniversalDependencies/tools/blob/master/udtools/README.md#the-official-udconll-u-validator

[3] https://github.com/isi-nlp/uroman

| Dataset / Model | Split | # Sentences | # Tokens |
|---|---|---|---|
| Griko | train | 575 | 6,992 |
| Griko | test | 64 | 783 |
| GUD (SMG) | train | 1,071 | 15,931 |

Table 1: Dataset statistics for the experimental setup. The Romanized GUD dataset corresponds to a script-transformed version of UD_Greek-GUD and therefore shares identical sentence and token counts. The Italian condition is based on Stanza's official pretrained combined Italian model; see the main text for further details.

**Italian Model.** To assess whether a contact-related language provides an advantage, we additionally evaluate cross-lingual transfer from Italian using Stanza's official pretrained Italian dependency parsing pipeline.

According to the Stanza documentation,[4] the default Italian package is a combined model trained on multiple UD Italian treebanks, namely ISDT, VIT, PoSTWITA, and TWITTIRO. Because this condition relies on a released pretrained package rather than on a model retrained under the same train/dev split regime as our other trainable conditions, it should be interpreted as a comparison baseline rather than as a fully controlled parallel training condition.

This condition is included as a convenient Italo-Romance comparison point rather than as a direct proxy for the local Italo-Romance varieties historically in contact with Griko.

### 3.2. Methods

**Training Configurations.** We evaluate four training configurations:

1. **Intra-dialectal transfer (Griko→Griko)**: The model is trained and evaluated exclusively on Griko data. This condition establishes an upper-bound baseline for parsing performance given available training data.
2. **Cross-dialectal transfer (SMG→Griko)**: The model is trained on SMG (UD_Greek-GUD) and evaluated on Griko. This condition tests the hypothesis that structural similarity within the MG ISO grouping enables knowledge transfer.
3. **Script-controlled transfer (Romanized SMG→Griko)**: The model is trained on romanized SMG and evaluated on Griko. This configuration controls for orthographic mismatch while preserving the SMG morphosyntactic specification.
4. **Cross-lingual (contact-related) transfer (Italian→Griko)**: The model is the official pretrained Stanza combined Italian model and is evaluated directly on the Griko test set. This condition tests whether a Romance baseline linked to the contact ecology of Griko offers a more informative transfer source than SMG.

All models are evaluated on the same held-out Griko test set.

[4] https://stanfordnlp.github.io/stanza/combined_models.html

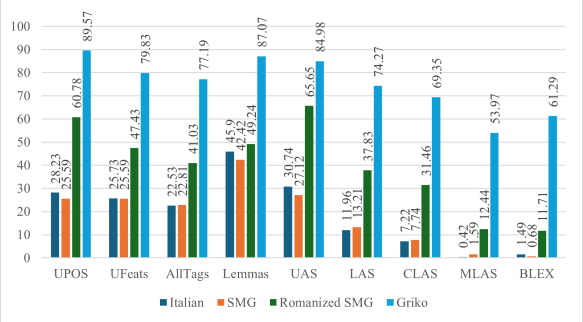


Figure 2: Dependency parsing performance (F1) on the Griko test set under four experimental conditions: intra-dialectal training, cross-dialectal transfer from SMG, script-controlled transfer using romanized SMG, and contact-related cross-lingual transfer from Italian.

**Model Architecture.** All experiments were conducted using the Stanza neural dependency parser (Qi et al., 2020). For SMG training, we use Greek-specific contextual embeddings (nlpaueb/bert-base-greek-uncased-v1) (Koutsikakis et al., 2020), which are appropriate for Greek-script data. For Griko and romanized SMG training, we use multilingual contextual embeddings (jhu-clsp/mmBERT-base) (Marone et al., 2025), which are better suited to Latin-script input and cross-dialectal generalization. For the Italian baseline, we use the official pretrained Stanza combined Italian model distributed with Stanza. Hyperparameters followed Stanza default configurations.

**Evaluation.** Parsing performance is evaluated using the official UD parsing scorer.[5] We report standard UD metrics, expressed as F1 scores.

## 4. Results

Figure 2 reports dependency parsing performance (F1 scores) across the four training configurations.

**Cross-dialectal Transfer (SMG→Griko).** When a model trained on SMG is evaluated on Griko, performance drops sharply across all metrics. Labeled Attachment Score (LAS) reaches only 13.21, while

[5] https://github.com/UniversalDependencies/tools/blob/master/udtools/README.md#the-official-ud-parsing-scorer

MLAS and BLEX—metrics sensitive to the interaction between morphology and syntax—approach zero (1.59 and 0.68, respectively).

Two factors may contribute to this degradation. First, there is substantial script distance between the varieties: SMG is written in the Greek alphabet, whereas Griko uses a Latin-based orthography. Because modern parsers rely on contextualized subword representations, orthographic mismatch introduces immediate divergence at the input level. Second, SMG and Griko differ in key morphosyntactic domains, including auxiliary inventories, voice marking, and progressive constructions. The near-total collapse of MLAS indicates that morphological feature prediction fails almost entirely under direct transfer, suggesting incompatibility at the level of feature systems and syntactic structuring rather than simple lexical mismatch.

Compared with Italian, SMG remains slightly better on LAS, CLAS, and MLAS, but weaker on several metrics before romanization. This contrast is important for our assumptions: relatedness to Greek does not automatically overcome script-induced and structural mismatch, while contact-related Romance similarity can help locally without yielding stronger syntactic transfer.

**Script-Controlled Transfer (Romanized SMG→Griko).** To disentangle orthographic effects from structural divergence, we train on a romanized version of SMG while preserving its morphosyntactic annotation.

Performance improves substantially relative to direct transfer: LAS increases from 13.21 to 37.83, UAS from 27.12 to 65.65, and UPOS to 60.78. These gains confirm that script discrepancy is a significant source of transfer degradation, as orthographic alignment facilitates lexical and subword generalization across varieties.

Romanized SMG also clearly outperforms both original-script SMG and Italian on every reported metric, showing that script normalization matters much more than the choice between those two unnormalized donor conditions. However, performance remains well below the intra-dialectal baseline. LAS reaches 37.83 compared to 74.27 in the Griko→Griko condition, and MLAS remains low at 12.44. The remaining gap indicates that structural divergence—not script difference alone—constitutes the primary limitation in cross-dialectal transfer.

**Contact-Related Cross-Lingual Transfer (Italian→Griko).** When the official pretrained Italian Stanza combined model is evaluated on Griko, performance is likewise poor overall: LAS reaches 11.96, CLAS 7.22, MLAS 0.42, and BLEX 1.49. At the same time, Italian slightly outperforms non-romanized SMG on UPOS, UFeats, Lemmas, UAS, and BLEX. This suggests that orthographic compatibility and contact-related lexical overlap can help some surface-oriented predictions; however, Italian does not yield better syntax-sensitive transfer than Greek.

**Intra-Dialectal Training (Griko→Griko).** When trained and evaluated exclusively on Griko data, the model achieves robust performance despite the relatively limited dataset size. This result shows that the Griko dataset forms a learnable and internally consistent system. The low performance observed under cross-dialectal transfer cannot therefore be attributed solely to data sparsity or annotation instability, but reflects genuine cross-dialectal divergence.

**Summary.** The four configurations reveal a consistent pattern:

1. Direct transfer from SMG to Griko yields near-complete failure.
2. Transfer from Italian is also poor overall, although it slightly outperforms non-romanized SMG on some metrics.
3. Script normalization substantially improves performance and clearly surpasses both original-script SMG and Italian, but leaves a large residual gap.
4. Intra-dialectal training produces strong results.

Taken together, these findings indicate that neither shared Greek ISO grouping nor a contact-related Romance baseline guarantees structural interoperability for dependency parsing. The fact that Italian can outperform non-romanized SMG on some metrics prior to romanization is consistent with the role of orthography and contact-related overlap. Crucially, the gap between Romanized SMG and Griko-to-Griko training shows that deeper morphosyntactic divergence remains the principal obstacle to transfer.

## 5. Annotation Constraints and Dialect-Specific Divergence

The experimental results indicate that cross-dialectal transfer between SMG and Griko remains highly limited, even after controlling for script differences. While romanization substantially improves performance, a large gap persists between cross-dialectal and intra-dialectal training. This residual gap points to deeper structural divergence rather than surface-level orthographic mismatch. In this section, we examine how this divergence interacts with UD's organizational principles, particularly its language-level validation framework.

### 5.1. Griko under ISO-Based Validation Coupling

Within UD, Griko is not assigned an independent ISO 639-3 code and is therefore classified under MG. As a result, it cannot be represented as a separate language-level treebank, but must conform to the MG specification, including its inventories of auxiliaries, morphological features, and language-specific validation checks enforced by the official UD validator.

We refer to this configuration as *ISO-based validation coupling*: varieties grouped under the same ISO code are validated against a shared grammatical specification, irrespective of internal structural divergence. Validation thus operates at the level of the language code, not at the level of individual varieties.

This arrangement has concrete consequences for annotation.

First, auxiliary inventories are defined and validated at the language level. Auxiliaries that exist in Griko but not in SMG cannot be treated as variety-specific elements without modifying the shared MG specification. Conversely, auxiliaries shared across varieties but realized in different scripts introduce additional complexity. In the Griko case, the use of the Latin alphabet prevents direct orthographic ambiguity; however, for dialects written in the Greek alphabet, lexical verbs could potentially be validated as auxiliaries without triggering language-level inconsistencies.

Second, morphological feature inventories are likewise shared across all MG treebanks. Griko maintains a systematic distinction among `Voice=Act`, `Voice=Mid`, and `Voice=Pass`, whereas SMG does not preserve a comparable tripartite opposition. Because feature values are validated at the language level, the distribution of Voice=Mid cannot be restricted to Griko alone. Dialect-specific grammatical contrasts therefore cannot be formally encoded without simultaneously altering the specification for all MG varieties.

In what follows, we illustrate these dynamics through three case studies—auxiliary formation, voice marking, and progressive constructions—which demonstrate how Griko reorganizes core morphosyntactic domains in ways that challenge language-level validation assumptions within UD.

### 5.2. Case Study I: Dialect-Specific Auxiliaries

Griko has developed auxiliaries that do not exist in SMG, notably *èrkome* and *ènna*.

***èrkome* as Eventive Passive Auxiliary** *Èrkome* functions as an eventive passive auxiliary in both present and past contexts, reflecting contact-induced grammaticalization from Italo-Romance.

(1) èrkete fameno
come.IND.PRS.3SG.MID eat.PTCP.PASS.NOM.M.SG
'He is (being) eaten.' (Baldissera, 2013, 43)

***ènna* as Modal Auxiliary** *Ènna* is a deontic modal auxiliary derived from a grammaticalized construction corresponding to SMG *ekhi na*. Unlike SMG, Griko has reanalyzed this construction into an invariable auxiliary. Because it encodes modality beyond TAM distinctions, it is annotated with `VerbType=Mod`.

(2) fseri ti
know.IND.PRS.2SG.ACT what

ènna kamin
AUX do.IND.PRS.PFV.2SG.ACT
'Do you know what you should do?'
(Tommasi, 1998, Tale 27)

These auxiliaries exemplify structural developments absent from SMG. Under ISO-based validation coupling, however, incorporating them requires modification of the shared MG auxiliary inventory, affecting all treebanks under the same code.

### 5.3. Case Study II: Middle vs. Passive Voice

Griko exhibits a systematic distinction between middle and passive constructions. Unlike SMG, which employs a unified non-active morphology, Griko restricts inherited non-active morphology to middle-type uses (reflexive, anticausative, unaccusative). Passive meaning is instead expressed periphrastically via *ime*/*èrkome* + perfect passive participle.

(3a) plènome
wash.IND.PRS.1SG.MID
'I wash myself.'

(3b) èrkome plimeno
come.IND.PRS.1SG.MID wash.PTCP.PERF.PASS.NOM.M.SG
'I am being/getting washed.'

In these periphrastic constructions, no dedicated `Voice=Pass` morphology appears on the auxiliaries themselves; passive interpretation is signaled by the participle. Although passive voice could in principle be inferred from interpretation, we encode it explicitly at the syntactic level by using the dependency relation `aux:pass`.

A further complication arises from the fact that, in older Griko and in some contemporary varieties, BE-perfect (4a) and BE-passive constructions (4b) are morphologically identical:

(4a) ime ertomeno / ertomeni
be.IND.PRS.1SG come.PTCP.PERF.PASS.NOM.M.SG / F.SG
'I have arrived (m/f).'

(4b) ime sfammeno / sfammeni
be.IND.PRS.1SG kill.PTCP.PERF.PASS.NOM.M.SG / F.SG
'I am killed (m/f).'

Following the approach adopted in UD_Greek-GUD, participles in *-menos* are annotated as passive verbal forms (Markantonatou et al., 2025). Since no morphological distinction differentiates perfect from passive in these constructions, the contrast is expressed syntactically: the dependency relation `aux:pass` is used in (4b) to mark passive interpretation explicitly. Annotating the participle in (4a) as a middle formation would remove the need for this syntactic distinction; however, such an analysis would depart from the generalizations established in GUD and compromise cross-treebank consistency.

In contemporary varieties of Griko, perfect participles tend to be invariable in *-mena*, whereas passive participles retain agreement.

### 5.4. Case Study III: Progressive and Pseudo-Coordination Constructions

Griko exhibits pseudo-coordination structures of the form V1 + coordinator + V2, conveying aspectual or discourse-related functions. Following SMG annotation practice, such constructions are annotated with `conj` (or `parataxis` when the coordinator is absent), and semantically bleached V1 verbs receive `VerbType=Mod`.

A particularly revealing case involves the progressive construction with *ste(o)* 'to stay', which appears in three stages of grammaticalization:

(5a) Inflected *steo* + ce + V2
ìstike ce piske
stay.IND.PST.IPFV.3SG CONJ fish.IND.PST.IPFV.3SG
'He was fishing.' (Tommasi, 1998, Tale 107)

(5b) Invariable *ste* + ce + V2
ste ce piske
stay.DEFAULT CONJ fish.IND.PST.IPFV.3SG
'He was fishing.'

(5c) Invariable *ste* + V2
ste piske
stay.DEFAULT fish.IND.PST.IPFV.3SG
'He was fishing.'

In cases (5a) and (5b), we maintain coordination analysis. In (5c), the absence of inflection and conjunction signals full grammaticalization, and *ste* is annotated as `AUX`, in accordance with UD guidelines.

A parallel pseudo-coordination pattern occurs with *pianno* 'to grab':

(6a) pianni ce lei
grab.IND.PRS.3SG CONJ say.IND.PRS.3SG
'He grabs and says.'

(6b) pianni èftase èssu-ti
grab.IND.PRS.3SG arrive.IND.PST.PFV.3SG inside-she.GEN.F.SG
'He grabbed and arrived at her place.'
(Tommasi, 1998, Tale 46)

Unlike *ste*, however, *pianni* does not encode aspectual information and is therefore analyzed as a lexical verb in coordination or parataxis.

### 5.5. Summary

Across auxiliary formation, voice marking, and progressive constructions, Griko exhibits systematic morphosyntactic reorganization shaped by prolonged contact and isolation. These are not superficial lexical differences but structural divergences affecting core grammatical domains.

The severe degradation observed in cross-dialectal transfer thus reflects genuine grammatical incompatibility. ISO-based validation coupling ensures administrative coherence within UD, but it does not guarantee structural compatibility across dialects. Incorporating Griko into UD therefore requires more than data integration; it demands reconsideration of how dialect-level grammatical divergence is accommodated within language-level validation architectures.

## 6. Discussion: Toward More Flexible Treatment of Dialects in UD

The Griko case points to a broader issue for UD: how can the framework adequately represent structurally coherent lects that are dialectal, may lack a standardized parent language, or do not have an ISO code of their own? Our results suggest that this is not a marginal technical issue, but a central methodological challenge for the representation of dialectal diversity in universal annotation frameworks.

UD does not need to abandon its universal layer to accommodate dialects more adequately, but it does require greater flexibility in how varieties are documented, validated, and organized. One practical step would be to introduce explicit *variety profiles* under a shared language node, coupled with layered validation in which a treebank is checked first against universal and language-level UD constraints and then against variety-sensitive rules at the treebank or dialect level. This would preserve comparability across related treebanks without imposing the same decisions on all treebanks grouped under a single language code. More explicit variety-level documentation would also help separate genealogical affiliation from annotation governance, especially since UD's higher-level classification draws heavily on WALS, whose language-centered family and genus groupings are too coarse to capture dialect-internal structural divergence.

The problem is even sharper for varieties that lack both a standardized parent language and an independent ISO code. In such cases, rep-

resentational adequacy should not depend exclusively on administrative languagehood. From this perspective, a more flexible organizational basis—potentially drawing on resources such as Glottolog (Hammarström et al., 2026), which is designed to catalogue documented *languoids* below the standardized-language level—could offer a more flexible and comprehensive classificatory basis than ISO- or WALS-based organization alone.

At the same time, UD already shows that some flexibility is possible: Cappadocian and Pomak are treated as separate languages rather than being folded into broader genealogical groupings, even where ISO coding is absent, as with Pomak. This suggests that the issue is not the lack of mechanisms, but the absence of a consistent principle for when a dialect receives its own profile versus remaining under a broader language-level specification.

More generally, the Griko case highlights a tension within UD between cross-dialectal comparability and dialect-specific adequacy. Our results show that this tension is not only descriptive but also computationally consequential: neither shared ISO grouping nor genealogical relatedness guarantees effective transfer, as illustrated by the contact-related Italo-Romance baseline slightly outperforming non-romanized SMG. A more flexible approach would allow dialectal lects to retain a shared universal annotation layer while receiving their own formal profiles when justified.

## 7. Conclusions and Future Work

This study examined the incorporation of Griko into the UD framework under the MG specification. Through controlled parsing experiments, we showed that transfer to Griko remains severely limited under both cross-dialectal and cross-lingual conditions. Direct transfer from SMG performs extremely poorly, even relative to an Italian baseline. Although script normalization substantially improves performance in the SMG condition, a large residual gap persists relative to intra-dialectal training. These results indicate that orthographic mismatch explains only part of the degradation, whereas deeper morphosyntactic divergence remains the principal obstacle to transfer. We further substantiate this interpretation through annotation-level analysis.

We argued that this divergence interacts with what we term *ISO-based validation coupling*: varieties grouped under the same ISO code are validated against a shared grammatical specification even when their internal structures differ substantially. In Griko, auxiliary inventories, voice distinctions, and progressive constructions reveal systematic reorganization in core grammatical domains. Because validation operates primarily at the language level, dialect-specific categories cannot be independently constrained without potentially affecting all MG treebanks. More broadly, the Griko case highlights the limits of UD's language-centric organization. Its higher-level classification, partly based on WALS, is useful for broad genealogical organization but too coarse to capture dialect-internal structural divergence.

More generally, these findings suggest that standard-centric modeling practices may overestimate cross-dialectal transferability and obscure dialect-specific grammatical systems. Griko is not computationally well approximated by genealogically related SMG; indeed, before romanization, the contact-related Romance baseline slightly surpasses SMG on some UD metrics. The strongest performance is achieved when the model is trained on Griko itself, indicating that the variety constitutes a coherent grammatical system in its own right. Taken together, these results show that genealogical relatedness and shared grouping do not guarantee syntactic compatibility for dependency parsing (Lin et al., 2019; Faisal and Anastasopoulos, 2022), particularly in contact settings and where script differences are involved (see Vakirtzian et al., 2024 and Tsoukala et al., 2026 for similar patterns in Greek ASR). Empirical evaluation, combined with annotation-level analysis, is therefore essential when extending universal annotation frameworks to structurally divergent dialects.

Incorporating dialects into UD thus requires more than simply adding new treebanks under an existing language label. It requires mechanisms that preserve the universal annotation layer while allowing dialect-specific structure to be documented and validated on its own terms. As discussed above, possible directions include explicit variety profiles, layered validation, clearer separation between genealogical affiliation and annotation governance, and more flexible organizational bases for lects without independent ISO codes.

Future work will expand the Griko treebank, compare Griko more systematically with other (MG) dialect treebanks, and test whether the observed pattern holds across additional parsing architectures and transfer settings. Further work within UD is also needed to explore how variety-sensitive validation mechanisms can be implemented in practice without sacrificing interoperability.

Ultimately, the Griko case shows that incorporating dialects into UD requires more than adding new data: it requires rethinking how a language-level validation architecture can accommodate structural diversity within genealogically related varieties, and how standard-centric design decisions shape the representation and computational modeling of dialect-specific grammatical systems.

## 8. Limitations

Several limitations should be acknowledged.

First, the Griko treebank is still under development and relatively small in size. Although intra-dialectal transfer yields robust performance, limited data may affect the stability of fine-grained morphosyntactic evaluation and the generalizability of absolute scores. Expanding the corpus will allow for more precise assessment of structural phenomena and model behavior.

Second, validation of the Griko data was restricted to format-level compliance using the official UD CoNLL-U validator. Language-specific morphological and syntactic validation rules were not fully enforced, as these are currently defined only for SMG under the shared ISO code. While this reflects the structural issue examined in the paper, it also means that some annotation-level inconsistencies may remain undetected.

Third, the experimental setup relies on a single parsing architecture (Stanza) and specific contextual embedding choices. Although these configurations are standard and appropriate for the respective scripts, results may vary under alternative modeling approaches. Future work should evaluate whether the observed transfer limitations persist across architectures and multilingual pretraining strategies.

Fourth, the Italian condition uses the official pretrained Stanza combined model rather than a model trained by us on a manually selected Italian treebank split. This makes the baseline informative as a practical contact-related comparison, but not fully controlled in the same way as the SMG, Romanized SMG, and Griko conditions.

Finally, this study focuses on a single non-standard variety shaped by long-term isolation and language contact. While the analysis highlights structural challenges arising from ISO-based validation coupling, further comparative work across additional dialects will be necessary to determine the broader generalizability of the findings.

These limitations do not undermine the central empirical observation—that structural divergence substantially constrains cross-dialectal transfer—but they delimit the scope within which the present conclusions should be interpreted.

## 9. Data Availability

All data, models, and scripts used in this study are publicly available at https://osf.io/acf2x/overview. The repository includes the Griko UD-style treebank, the original UD_Greek-GUD dataset and its romanized version used in the script-controlled experiments, the corresponding Python romanization script, and the trained Stanza dependency parsing models for all experimental configurations. For the Italian baseline, we report the evaluation scores obtained with the official pretrained Stanza combined Italian model.

## 10. Acknowledgements

This work has been partially supported by project MIS 5154714 of the National Recovery and Resilience Plan Greece 2.0 funded by the European Union under the NextGenerationEU Program.

We are especially grateful to the Griko speakers for their time, generosity, and invaluable help throughout this work.

## 11. Bibliographical References

Antonios Anastasopoulos, Marika Lekakou, Josep Quer, Eleni Zimianiti, Justin DeBenedetto, and David Chiang. 2018. Part-of-Speech Tagging on an Endangered Language: a Parallel Griko-Italian Resource. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2529–2539, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Valeria Baldissera. 2013. *Il dialetto grico del Salento: elementi balcanici e contatto linguistico*. Ph.D. dissertation, Università Ca' Foscari Venezia.

Verena Blaschke, Barbara Kovačić, Siyao Peng, Hinrich Schütze, and Barbara Plank. 2024. MaiBaam: A Multi-Dialectal Bavarian Universal Dependency Treebank. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10921–10938, Torino, Italia. ELRA and ICCL.

Marcely Zanon Boito, Antonios Anastasopoulos, Marika Lekakou, Aline Villavicencio, and Laurent Besacier. 2018. A small Griko-Italian speech translation corpus. *CoRR*, abs/1807.10740. ArXiv: 1807.10740.

Stavros Bompolas, Stella Markantonatou, Angela Ralli, and Antonios Anastasopoulos. 2025. Crossing dialectal boundaries: Building a treebank for the dialect of lesbos through knowledge transfer from standard Modern Greek. In *Proceedings of the Eighth Workshop on Universal Dependencies (UDW, SyntaxFest 2025)*, pages 39–51, Ljubljana, Slovenia. Association for Computational Linguistics.

Richard MacGillivray Dawkins. 1916. *Modern Greek in Asia Minor: a study of the dialects of Sílli,*

*Cappadocia and Phárasa with grammar, texts, translations and glossary*. Cambridge University Press, Cambridge.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online (v2020.4)*. Zenodo.

Fahim Faisal and Antonios Anastasopoulos. 2022. Phylogeny-Inspired Adaptation of Multilingual Models to New Languages. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 434–452, Online only. Association for Computational Linguistics.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2026. glottolog/glottolog: Glottolog database 5.3.

Ulf Hermjakob, Jonathan May, and Kevin Knight. 2018. Out-of-the-box universal Romanization tool uroman. In *Proceedings of ACL 2018, System Demonstrations*, pages 13–18, Melbourne, Australia. Association for Computational Linguistics.

John Koutsikakis, Ilias Chalkidis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2020. GREEK-BERT: The Greeks visiting Sesame Street. In *11th Hellenic Conference on Artificial Intelligence*, SETN 2020, pages 110–117, New York, NY, USA. Association for Computing Machinery. Event-place: Athens, Greece.

Andre Kåsen, Kristin Hagen, Anders Nøklestad, Joel Priestly, Per Erik Solberg, and Dag Trygve Truslew Haug. 2022. The Norwegian Dialect Corpus Treebank. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4827–4832, Marseille, France. European Language Resources Association.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing Transfer Languages for Cross-Lingual Learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy. Association for Computational Linguistics.

Mohamed Maamouri, Ann Bies, Seth Kulick, Michael Ciul, Nizar Habash, and Ramy Eskander. 2014. Developing an Egyptian Arabic Treebank: Impact of Dialectal Morphology on Annotation and Tool Development. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2348–2354, Reykjavik, Iceland. European Language Resources Association (ELRA).

Io Manolessou. 2005. The Greek Dialects of Southern Italy: An Overview. *Cambridge Papers in Modern Greek*, 13:103–125.

Stella Markantonatou, Vivian Stamou, Stavros Bompolas, Katerina Anastasopoulou, Irianna Linardaki Vasileiadi, Konstantinos Diamantopoulos, Yannis Kazos, and Antonios Anastasopoulos. 2025. VMWE identification with models trained on GUD (a UDv.2 treebank of Standard Modern Greek). In *Proceedings of the 21st Workshop on Multiword Expressions (MWE 2025)*, pages 14–20, Albuquerque, New Mexico, U.S.A. Association for Computational Linguistics.

Stella Markantonatou, Nicolaos Th. Constantinides, Vivian Stamou, Vasileios Arampatzakis, Panagiotis G. Krimpas, and George Pavlidis. 2023. Methodological issues regarding the semi-automatic UD treebank creation of under-resourced languages: the case of Pomak. In *Proceedings of the Sixth Workshop on Universal Dependencies (UDW, GURT/SyntaxFest 2023)*, pages 27–35, Washington, D.C. Association for Computational Linguistics.

Marc Marone, Orion Weller, William Fleshman, Eugene Yang, Dawn Lawrie, and Benjamin Van Durme. 2025. mmBERT: A Modern Multilingual Encoder with Annealed Language Learning. _eprint: 2509.06888.

Aleksandra Miletic, Myriam Bras, Marianne Vergez-Couret, Louise Esher, Clamença Poujade, and Jean Sibille. 2020. A Four-Dialect Treebank for Occitan: Building Process and Parsing Experiments. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 140–149, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Christopher Moseley. 2012. *The UNESCO atlas of the world's languages in danger: context and process*. World Oral Literature Project. OCLC: 811994673.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis

Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Manuela Pellegrino. 2021. *Greek Language, Italian Landscape: Griko and the Re-storying of a Linguistic Minority*. Harvard University Press / Center for Hellenic Studies.

Prokopis Prokopidis and Haris Papageorgiou. 2017. Universal Dependencies for Greek. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 102–106, Gothenburg, Sweden. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Alan Ramponi. 2024. Language Varieties of Italy: Technology Challenges and Opportunities. *Transactions of the Association for Computational Linguistics*, 12:19–38.

Shruti Rijhwani, Antonios Anastasopoulos, and Graham Neubig. 2020. OCR Post Correction for Endangered Language Texts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5931–5942, Online. Association for Computational Linguistics.

Konstantinos Sampanis and Prokopis Prokopidis. 2021. Asia Minor Greek in Contact (AMGiC): Towards a dialectal Treebank comprising contact-induced grammatical changes. In *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)*, pages 86–95, Sofia, Bulgaria. Association for Computational Linguistics.

Salvatore Tommasi. 1998. *Io' mia forà… Fiabe e Racconti della Grecìa Salentina. Dai quaderni (1883–1912) di Vito Domenico Palumbo*. Ghetonia, Galatina, Lecce.

Chara Tsoukala, Stavros Bompolas, Antigoni Margariti, Konstantina Panagiotou, Maria Elisavet Plaiti, Nefeli Tzanakaki, Petros Karatsareas, Angela Ralli, Antonios Anastasopoulos, and Stella Markantonatou. 2026. Extending ASR Evaluation Resources for Modern Greek Dialects. In *Proceedings of the 13th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 210–222, Rabat, Morocco. Association for Computational Linguistics.

Socrates Vakirtzian, Vivian Stamou, Yannis Kazos, and Stella Markantonatou. 2025. Dialectal treebanks and their relation with the standard variety: The case of East Cretan and Standard Modern Greek. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 776–784, Tallinn, Estonia. University of Tartu Library.

Socrates Vakirtzian, Chara Tsoukala, Stavros Bompolas, Katerina Mouzou, Vivian Stamou, Georgios Paraskevopoulos, Antonios Dimakis, Stella Markantonatou, Angela Ralli, and Antonios Anastasopoulos. 2024. Speech Recognition for Greek Dialects: A Challenging Benchmark. In *Interspeech 2024*, pages 3974–3978. ISCA.

Eleni Vligouridou, Inessa Iliadou, and Çağrı Çöltekin. 2024. A Treebank of Asia Minor Greek. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1715–1721, Torino, Italia. ELRA and ICCL.

Ruoyu Xie and Antonios Anastasopoulos. 2023. Noisy Parallel Data Alignment. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1501–1513, Dubrovnik, Croatia. Association for Computational Linguistics.

Lilja Øvrelid, Andre Kåsen, Kristin Hagen, Anders Nøklestad, Per Erik Solberg, and Janne Bondi Johannessen. 2018. The LIA Treebank of Spoken Norwegian Dialects. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).